Journal of Research on Educational Effectiveness, 8: 280–302, 2015

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2014.911396

Routledge
Taylor & Francis Group

Covariate Balance in Bayesian Propensity Score Approaches for Observational Studies

Jianshen Chen and David Kaplan
University of Wisconsin–Madison, Madison, Wisconsin, USA

Abstract: Bayesian alternatives to frequentist propensity score approaches have recently been proposed. However, few studies have investigated their covariate balancing properties. This article compares a recently developed two-step Bayesian propensity score approach to the frequentist approach with respect to covariate balance. The effects of different priors on covariate balance are evaluated and the differences between frequentist and Bayesian covariate balance are discussed. Results of the case study reveal that both the Bayesian and frequentist propensity score approaches achieve good covariate balance. The frequentist propensity score approach performs slightly better on covariate balance for stratification and weighting methods, whereas the two-step Bayesian approach offers slightly better covariate balance in the optimal full matching method. Results of a comprehensive simulation study reveal that accuracy and precision of prior information on propensity score model parameters do not greatly influence balance performance. Results of the simulation study also show that overall, the optimal full matching method provides the best covariate balance and treatment effect estimates compared to the stratification and weighting methods. A unique feature of covariate balance within Bayesian propensity score analysis is that we can obtain a distribution of balance indices in addition to the point estimates so that the variation in balance indices can be naturally captured to assist in covariate balance checking.

Keywords: Propensity scores, covariate balance, Bayesian statistics

INTRODUCTION

It is well-established that in randomized experiments, individuals are assigned to treatment conditions with a known probability. By contrast, in observational studies, individuals self-select into treatment conditions on the basis of an unknown mechanism. Thus, the selection process is often highly nonrandom, introducing selection bias that may result in highly unbalanced covariates (i.e., greatly different covariate distributions in the treatment group and control group) and thus severely weakening causal inferences. The best that can be hoped for is that the investigator has obtained measurable and reliable covariates that relate to the selection mechanism and the potential outcomes, recognizing that unobservable covariates might still be operating to bias treatment effect estimates. Establishing balance between the treatment and control groups on observable covariates is thus essential for obtaining unbiased treatment effect estimates.

Address correspondence to Jianshen Chen, 660 Rosedale Road, Princeton, NJ 08534, USA. E-mail: jchen006@ets.org

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uree.

In a classic article, Rosenbaum and Rubin (1983) proposed propensity score analysis as a practical tool for reducing selection bias through balancing on measured covariates, where the propensity score is a scalar function of covariates so that subjects who match on their propensity scores can be treated as having similar covariate background. A variety of propensity score techniques have been developed for both the estimation and the application of the propensity score. Models for estimating the propensity score equation have included, for example, parametric logit regression with chosen interaction and polynomial terms (e.g., Dehejia & Wahba, 1999; Hirano & Imbens, 2001), and nonparametric generalized boosting modeling (McCaffrey, Ridgeway, & Morral, 2004). Methods for estimating the treatment effect while accounting for the propensity score include stratification, weighting, matching, and regression adjustment. Each of these propensity score techniques have a common goal, which is to achieve balanced covariate distributions across the different treatment conditions. Covariate balance after propensity score adjustment, therefore, is often evaluated to assess the effectiveness of propensity score techniques (see, e.g., Harder, Stuart & Anthony, 2010).

There have been numerous research studies examining the adequacy of covariate balance. For example, Rubin (2001) suggested three criteria for gauging balance between the treatment and control group, namely, small mean differences on propensity scores between the treatment and control group, similar variances of the propensity scores between treatment and control conditions, and similar residual variances of the covariates after the propensity score adjustment (the variance ratio should be close to one) in the two groups. Later, Austin and Mamdani (2006) demonstrated various indices to evaluate covariate balance, such as standardized differences and nonparametric density estimates. Among approaches for assessing covariate balance, statistical hypothesis testing has often been utilized in practice. However, there are two major complaints shared by Imai, King, and Stuart (2008) and Austin (2008) with regard to the validity of balance testing: (a) covariate balance is a sample property, and thus hypothesis testing is not relevant, and (b) sample-size reduction in matched samples can lead to less significant imbalance results even if the absolute imbalances remain constant. Agreement surrounding these concerns, though, has not been fully reached (Hansen, 2008; Hill, 2008; Stuart, 2008). In this context, a Bayesian perspective is more appealing because it can naturally avoid both problems and is the subject of this article.

Propensity score analysis has been mainly developed within the frequentist framework of statistics. An alternative Bayesian perspective for propensity score analysis was originally advocated by Rubin (1985). McCandless, Gustafson, and Austin (2009) first provided a practical Bayesian propensity score approach that stratifies individuals on the quintiles of the estimated propensity score, treating the propensity score as a latent variable and jointly estimating the propensity score and outcome models. McCandless, Gustafson, Austin, and Levy (2009) examined the covariate balance of their joint modeling approach via a case study of the effectiveness of beta-blocker therapy in heart failure patients. Their Bayesian stratification approach with simultaneous estimation of the propensity score and treatment effect was shown to provide poorer balance than the conventional frequentist counterpart, but it did reduce the association between the covariates and outcome and thus reduced the confounding effects of unbalanced covariates.

A concern surrounding the work of McCandless, Gustafson, and Austin (2009) centers on the joint modeling of the propensity score and outcome equations. Specifically, the propensity score should only be determined by covariates measured prior to treatment implementation and not influenced by the outcome measured after treatment assignment Steiner and Cook (2013). To address this problem, Kaplan and Chen (2012) developed a

two-step Bayesian propensity score approach, fitting the propensity score model and outcome model separately and examining its performance in regard to treatment effect and variance estimates via propensity score stratification, weighting, and optimal full matching methods. However, the balance properties of the two-step Bayesian propensity score approach have not yet been evaluated.

This article investigates the balance properties of the two-step Bayesian propensity approach (Kaplan & Chen, 2012) via a case study and a simulation study and demonstrates a way of assessing covariate balance within the Bayesian framework. The link between covariate balance and treatment estimation is also examined and discussed. The uniqueness of evaluating covariate balance in Bayesian propensity score approaches is that the uncertainty of balance indices can be naturally accounted for through the posterior distribution of propensity score model parameters. The posterior probability intervals (i.e., credible intervals) of balance indices can be acquired to judge the adequacy of covariate balance beyond simply point estimates. Concern (a), just mentioned with regard to balance tests, is irrelevant here because a Bayesian perspective draws inferences on the model parameters conditional on the sample and does not involve a hypothetical population from which the sample is derived. Concern (b) with regard to the impact of sample-size reduction accompanying matching on balance tests does not apply either because for Bayesian propensity score matching, uncertainty in the balance indices is captured by posterior distributions based on all the subjects, unlike frequentist hypothesis testing that examines significance of balance indices in matched samples with reduced sample sizes.

For the purpose of completeness, this article first briefly introduces the framework of propensity score analysis and its implementation. Then the following section presents Bayesian propensity score analysis and provides a comprehensive review of the Kaplan and Chen's (2012) two-step Bayesian propensity score approach. Next, the study design as well as results of a case study are presented and discussed, followed by the design and results of a simulation study. Finally, the conclusion section summarizes the article and discusses the pros and cons of covariate balance within Bayesian propensity score approaches.

PROPENSITY SCORE APPROACH AND ITS IMPLEMENTATION

Throughout this article, we consider the simple case of a two-group problem, where individuals self-select into a treatment group or a control group. The propensity score is defined to be the conditional probability that a participant selects into the treatment group given a vector of covariates measured prior to treatment selection. Following the general notation of the Rubin causal model (Rubin, 1974), let Y_{1i} be the potential outcome for individual i if the individual receives the treatment and Y_{0i} be the potential outcome for individual i if the individual does not receive the treatment. The treatment indicator is denoted as T, where for individual i, $T_i = 1$ if the individual was exposed to the treatment, and $T_i = 0$ if the individual was not exposed to the treatment. The propensity score of individual i given a vector of observed covariates x can be expressed as $e_i(x_i) = P(T_i = 1|x_i)$. For the treatment effect estimation based on the population of individuals, there are two causal estimands that may be of interest. One is the *average treatment effect* (ATE), defined as

$$\gamma_{ATE} = E(Y_{1i}) - E(Y_{0i}), \tag{1}$$

which is the mean difference between a unit assigned to the treatment condition and a unit from the same population but assigned to the control condition. The other causal estimand

is the *average treatment effect on the treated* (ATT), defined as

$$\gamma_{ATT} = E(Y_{1i}|T_i = 1) - E(Y_{0i}|T_i = 1), \tag{2}$$

which assumes that an individual is assigned to the treatment group and the question concerns the outcome of that individual had he or she been assigned to the control group. In this article, we focus on estimates of the ATE.

All propensity score techniques rely on the assumption of *strong ignorability* (Rosenbaum & Rubin, 1983), which states that the potential outcomes are assumed to be independent of treatment assignment given observed covariates. If strong ignorability does not hold, then hidden biases due to unobserved covariates may exist. Under strong ignorability, participants with the same propensity score are expected to have the same distribution on the set of covariates, here referred to as "covariate balance." Assuming strong ignorability and that covariate balance obtains, estimates of treatment effects from nonrandomized studies approximate those that would have been obtained if a randomized study was conducted. In practice, however, the true propensity score is unknown, and instead propensity scores are estimated by fitting a model such as a logistic regression model of the treatment assignment on the measured covariates. Then, based on the estimated propensity scores, different techniques, including stratification (Rosenbaum & Rubin, 1983, 1984), inverse-propensity weighting (Rosenbaum, 1987; Wooldridge, 2002), matching (Hansen, 2004; Rosenbaum, 1989), and regression estimation (Rubin, 1979), are utilized in the outcome model to obtain the estimate of the causal effect of interest.

As noted, there are several common methods for implementing a propensity score adjustment. The propensity score stratification method (Rosenbaum & Rubin, 1983) incorporates the idea of Cochran (1968) that five subclasses can remove as high as 90% of the bias due to the subclassifying covariate. The stratification method first sorts all the participants by their estimated propensity scores $\hat{e}(x)$ and then stratifies them into five or more strata at the quintiles or certain quantiles of the propensity score distribution. Estimates of the treatment effect within each stratum are calculated and the overall causal effect is obtained by averaging over five or more within-strata treatment effects. This method has seen increased popularity in educational, psychological, and medical research (e.g., Ayanian, Landrum, Guadagnoli, & Gaccione, 2002; Leow, Marcus, Zanutto, & Boruch, 2004; Swanson et al., 2007).

The inverse propensity score weighting approach is based on the Horvitz–Thompson estimator Horvitz and Thompson (1952) from the survey sampling literature. The inverse-propensity weighting approach weights participants in treatment group and control group by $1/\hat{e}(x)$ and $1/(1 - \hat{e}(x))$, respectively, to balance these groups on covariates.[1] The details of this approach can be found in Hirano and Imbens (2001); Hirano, Imbens, and Ridder (2003); and Lunceford and Davidian (2004).

Another technique, known as propensity score matching, matches participants in the treatment group to the control group on the estimated propensity score. Various matching methods have been explored in the context of propensity score analysis such as nearest-neighbor matching, caliper matching, Mahalanobis metric matching, and optimal matching (see, e.g., An, 2010; Hansen, 2004). Matching types can be one-to-one, one-to-many, many-many matching, and so on. The optimal matching method is "optimal" in the sense that each

[1]This particular weight yields the average treatment effect. One could weight by $T + (1 - T)\frac{\hat{e}(x)}{1-\hat{e}(x)}$, where $T = 1$ if the individual is assigned to the treatment group $T = 0$, if not. This weight provides an estimate of the average treatment effect on the treated.

match can be revisited to minimize the total distance between all the matches rather than removing the matched subjects from further consideration in the greedy matching method (Rosenbaum, 1989). Examples of the matching approach can be found in Foster (2003); Dehejia and Wahba (2002); and Guo, Barth, and Gibbons (2006).

Finally, the regression adjustment approach or the covariate-adjustment approach directly includes the estimated propensity score linearly (and possibly nonlinearly) in the outcome equation (Kang & Schafer, 2007; Schafer & Kang, 2008). The regression adjustment approach is easy to implement but relies more on the correct specification of the outcome model compared to other propensity score methods. Applications of this approach can be found in Kurth et al. (2006) and Shadish et al. (2008).

BAYESIAN PROPENSITY SCORE APPROACH

Over the past two decades, significant advances have been made in the area of Bayesian statistical inference, owing mostly to computational developments and readily available software (e.g., Gilks, Richardson, & Spiegelhalter, 1996). These computational advances have led to increased applications of Bayesian methods to problems in the social and behavioral sciences. As is well known, the Bayesian perspective begins by specifying a model for an outcome of interest, elicits prior distributions for all model parameters, and obtains the joint posterior distribution of the model parameters given the data via Bayes' theorem and some chosen computing algorithm, such as the Markov chain Monte Carlo (MCMC) methods including the Metropolis-Hastings algorithm, the Gibbs sampler, and so on. For a general review specific to the social and behavioral sciences, see Kaplan (2014).

Rubin (1985) argued that a Bayesian approach to propensity score analysis should be of great interest to the applied Bayesian analyst, and yet propensity score estimation within the Bayesian framework was not addressed until relatively recently. Hoshino (2008) developed a quasi-Bayesian estimation method for general parametric models, such as latent variable models, and developed an MCMC algorithm to estimate the propensity score. McCandless, Gustafson, and Austin (2009) first provided a practical Bayesian approach to propensity score stratification, estimating the propensity score and the treatment effect and sampling from the joint posterior distribution of model parameters via an MCMC algorithm. The marginal posterior probability of the treatment effect can then be obtained based on the joint posterior distribution. Similar to the McCandless, Gustafson, and Austin (2009) study, An (2010) presented a Bayesian approach that jointly models both the propensity score equation and outcome equation at the same time and extended this one-step Bayesian approach to propensity score regression and single nearest neighbor matching methods.

A consequence of the Bayesian joint modeling procedure utilized by McCandless, Gustafson, and Austin (2009) and An (2010) is that the propensity score estimates may be affected by the outcome variable that are observed after treatment assignment, resulting in biased propensity score estimation. In fact, Zigler et al. (2013) assessed the joint estimation of the Bayesian propensity score and the treatment effect and found that the feedback between propensity score model and outcome model can lead to poor treatment effect estimates. This model feedback is especially problematic if the relationship between the outcome and the propensity score is misspecified (McCandless, Douglas, Evans, & Smeeth, 2010). To solve this problem, McCandless, Douglas, Evans, and Smeeth (2010) utilized an approximate Bayesian technique introduced by Lunn et al. (2009) for preventing undesirable feedback between propensity score model and outcome model components. Specifically, McCandless, Douglas,

Evans, and Smeeth (2010) included the posterior distribution of the propensity score parameters as covariate input in the outcome model so that the flow of information between the propensity score and the outcome is restricted. This so-called *sequential Bayesian propensity score analysis* yields treatment effect estimates that are comparable to estimates obtained from frequentist propensity score analysis. Nevertheless, as McCandless, Douglas, Evans, and Smeeth (2010) pointed out, their method is only approximately Bayesian and also encounters the difficulty that the Markov chain is not guaranteed to converge.

The Two-Step Bayesian Propensity Score Approach

To maintain a fully Bayesian specification while overcoming the conceptual and practical difficulties of the joint modeling methods of McCandless, Gustafson, and Austin (2009) and An (2010), a two-step Bayesian propensity score approach was recently developed by Kaplan and Chen (2012) that can incorporate prior information on the model parameters of both the propensity score equation and outcome model equation. Consistent with Bayesian theory (see, e.g., De Finetti, 1974), specifying prior distributions on the model parameters is a natural way to quantify uncertainty—here in both the propensity score and outcome equations.

In the Kaplan and Chen (2012) two-step Bayesian propensity score approach (BPSA), the propensity score model is fit in the first step, specified as the following logit model.

$$Log\left(\frac{e(x)}{1-e(x)}\right) = \alpha + \beta' x, \tag{3}$$

where α is the intercept, β refers to the slope, and x represents a design matrix of chosen covariates. For BPSA, the R package *MCMClogit* (Martin, Quinn, & Park, 2010) was utilized to sample from the posterior distribution of α and β using a random walk Metropolis algorithm. Other R packages such as *arm* (Gelman et al., 2011) can also be applied to draw posterior samples from the posterior distribution. After the posterior samples of the propensity score are obtained, a Bayesian linear model is fit in the second step as the outcome model to obtain the posterior draws of the treatment effect via various propensity score methods such as stratification, weighting, optimal matching, and regression adjustment. We refer to the Bayesian linear model in the second step as the *Bayesian outcome model*. An illustration of the two-step Bayesian propensity score approach is shown in Figure 1.

For illustration purposes, consider a posterior sampling procedure of a chosen Bayesian logit model with 1, 000 iterations and a thinning interval of 1. Then for each observation, there are $m = 1,000$ posterior propensity scores $\hat{e}(x)$ calculated from the posterior samples of propensity score model parameters α and β as follows:

$$\hat{e}(x) = \frac{exp(\alpha + \beta' x)}{1 + exp(\alpha + \beta' x)}. \tag{4}$$

Based on each posterior propensity score, there are $J = 1,000$ posterior draws of the treatment effect generated from the posterior distribution of γ ($i = 1, \ldots, m$, $j = 1, \ldots, J$), where γ is the treatment effect. For notational simplicity, let η denote the vector of propensity score model parameters. Kaplan and Chen (2012) then provide the following

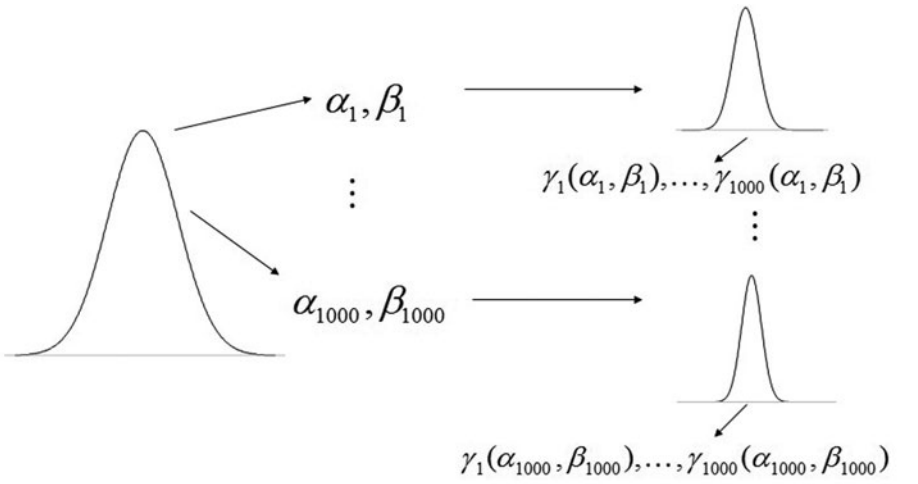

Figure 1. Illustration of the two-step Bayesian propensity score approach.

treatment effect estimator:

$$E(\gamma \mid x, y, T) = m^{-1} J^{-1} \sum_{i=1}^{m} \sum_{j=1}^{J} \gamma_j(\eta_i), \tag{5}$$

where $J^{-1} \sum_{j=1}^{J} \gamma_j(\eta_i)$ is the posterior sample mean of treatment effect γ in the Bayesian outcome model based on the *i*th set of propensity score model parameters η_i and then this posterior sample mean is averaged over *m* sets of posterior propensity scores. The posterior variance of γ is based on the total variance formula as follows:

$$Var(\gamma \mid x, y, T) = m^{-1} \sum_{i=1}^{m} \sigma^2_{\gamma(\eta_i)} + (m-1)^{-1} \sum_{i=1}^{m} \left\{ \mu_{\gamma(\eta_i)} - m^{-1} \sum_{i=1}^{m} \mu_{\gamma(\eta_i)} \right\}^2, \tag{6}$$

where

$$\sigma^2_{\gamma(\eta_i)} = (J-1)^{-1} \sum_{j=1}^{J} \left\{ \gamma_j(\eta_i) - J^{-1} \sum_{j=1}^{J} \gamma_j(\eta_i) \right\}^2, \tag{7}$$

is the posterior sample variance of γ in the Bayesian outcome model under the i^{th} set of propensity scores and

$$\mu_{\gamma(\eta_i)} = J^{-1} \sum_{j=1}^{J} \gamma_j(\eta_i), \tag{8}$$

is the posterior sample mean of γ in the same Bayesian outcome model. Note that Equation 6 captures two sources of variation. The first source of variation is the average of the

posterior variances of treatment effect γ across the posterior samples of propensity scores, represented by the first part of the right hand side of Equation 6, and the second source of variation comes from the variance of the posterior means of treatment effect γ across the posterior samples of propensity scores, obtained by the second part of the right of hand side of Equation 6.

Utilizing the aforementioned estimators, Kaplan and Chen (2012) conducted three comprehensive simulation studies as well as a small case study comparing frequentist propensity score analysis with the two-step Bayesian alternative focusing on the treatment effect and variance estimates. The effects of different sample sizes, true treatment effects and choice of priors on treatment effect and variance estimates were also evaluated. Consistent with Bayesian theory, Kaplan and Chen's findings showed that when no prior information is available, specifying a larger prior variance of treatment effect is desirable to obtain estimates similar to frequentist results but with more accurate intervals; when accurate prior information with regard to the treatment effect is attainable, specifying a smaller prior variance is preferable in order to obtain more precise treatment effect estimates. For the case of small sample size, the Bayesian approach shows slight superiority in the estimation of the treatment effect compared to the frequentist counterpart.

Kaplan and Chen (2012) studied the treatment effect and variance estimates of the two-step Bayesian propensity score approach, but the covariate balance of their approach has not been examined. Therefore, this article investigates covariate balance of the newly developed two-step full Bayesian propensity score approach under stratification, weighting, and optimal full matching methods via a case study and a real-data-based simulation study. We demonstrate a practical way of assessing the balance properties of the Bayesian propensity score approach via the point estimates as well as the posterior distribution of balance indices and also link the covariate balance results with the accuracy of treatment effect estimates. In addition, we discuss the unique features of the Bayesian approach in terms of covariate balance.

DESIGN OF THE CASE STUDY

A case study is conducted to examine properties of the two-step Bayesian propensity score approach. The data for this case study come from the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K) of 1998 (NCES, 2001). The sampled children attended either full-day or part-day kindergarten programs and have diverse socioeconomic and racial/ethnic backgrounds. Also, a number of variables assessing early childhood (pre-K) experiences were collected in the ECLS-K; thus propensity score approaches can be fruitfully applied. This article investigates the treatment effect of full-day versus part-day kindergarten attendance on children's reading achievement at the end of 1998 fall kindergarten.

A sample of 1,000 children were randomly selected proportional to the number of children in full-day or part-day kindergarten in the population. This resulted in 538 children in full-day programs and 462 children in part-day programs. Fourteen covariates were chosen for estimating the propensity scores, including gender, race, mother's employment status, child's age at kindergarten entry, child's age at first nonparental care, primary type of nonparental care, language spoken to child by parent, number of siblings, family composition, mother's employment between child's birth and kindergarten, number of nonparental care arrangements pre-K, social economic status, parent's expectation of child's degree, and how often parent reads to child. All analyses utilized various packages within

the R software environment (R Development Core Team 2011; described next), and missing data were handled via multiple imputations using the R program *mice* package (Van Buuren & Groothuis-Oudshoorn, 2011). Default settings in the *mice* program were used.

The procedure for assessing covariate balance is as follows. First, propensity scores are estimated via a Bayesian logit model. We use the "bayesglm" function of the *arm* package (Gelman et al., 2011) in R that adopts an approximate EM algorithm to update the parameters at each step to obtain the posterior samples of parameters in the Bayesian logit model. To handle problematic covariates, the "bayesglm" package uses a default Student's *t* prior for all regression coefficients. The posterior sampling runs 1,000 iterations, and thus we obtain 1,000 sets of propensity score estimates in the posterior sample. We then apply the estimated propensity scores to balance on covariates via propensity score stratification, weighting and optimal full matching methods. After having the matched strata/subgroups based on estimated propensity scores for stratification and optimal full matching, stratum frequency based weights are utilized to calculate weighted averages of covariates across matched strata/subgroups for the treatment group and the control group separately. Then the balance indices are obtained through taking the difference or the ratio of the weighted averages between the treatment group and the control group. In terms of the weighting method, inverse propensity score weights are used in the same way as stratum weights above to obtain estimated balance indices. For the optimal full matching method, the "optmatch" package in R (Hansen & Klopfer, 2006) is adopted.

For both the case study and the simulation study described next, the balance of continuous covariates and categorical covariates is evaluated separately. The balance indices used in this study are the standardized mean/proportion difference (Cohen's *d*; Cohen, 1988) and variance ratio for each continuous covariate/each level of categorical covariates between treatment group and control group. Specifically, the standardized mean difference for a continuous covariate is obtained by

$$B_1 = (\bar{x_t} - \bar{x_c})/\sqrt{\left(s_t^2 + s_c^2\right)/2}, \tag{9}$$

where $\bar{x_t}$ and $\bar{x_c}$ are the sample mean of each covariate in treatment and control groups, respectively, and s_t^2 and s_c^2 are corresponding sample variances. The variance ratio for a continuous covariate, R_1, is defined as s_t^2/s_c^2. All the categorical covariates are dummy coded. Then for each categorical level, we evaluate the standardized difference in proportions between different treatment conditions, consistent with Harder, Stuart and Anthony (2010). The standardized proportion difference is calculated by

$$B_2 = (\hat{p}_t - \hat{p}_c)/\sqrt{[\hat{p}_t(1 - \hat{p}_t) + \hat{p}_c(1 - \hat{p}_c)]/2}, \tag{10}$$

where $\hat{p}_t$ and $\hat{p}_c$ are proportions of participants in the treatment group and control group, respectively, for a specific level of categorical covariates. The variance ratio for a certain categorical level, R_2, is calculated by $\hat{p}_t(1 - \hat{p}_t)/\hat{p}_c(1 - \hat{p}_c)$.

In addition to the aforementioned point estimates, the Bayesian propensity score approach provides the 95% posterior probability intervals (*PPI*) of the standardized mean/proportion difference and the variance ratio based on the posterior distribution of the propensity score parameters. For each set of estimated propensity scores in the posterior sample, we can obtain a point estimate of each balance index. Because we have 1,000 sets of estimated propensity scores, a distribution of the balance index with 1,000 points can be obtained, and we extract the mean of this posterior distribution as the point estimate of the covariate balance as well as the 2.5 and 97.5 percentiles to form the corresponding 95%

Table 1. Average absolute standardized mean/proportion difference (Cohen's d), average variance ratio between treatment and control group across all covariates and treatment effect and standard error estimates (Trt. (SE)) in the case study

Method		Cohen's d	Variance Ratio	Trt. (SE)
Unadjusted	Bayesian	0.11	1.26	1.74 (0.77)
Stratification	Frequentist	0.02	1.07	2.00 (0.76)
	Bayesian (Noninformative)	0.03	1.08	2.05 (0.82)
Weighting	Frequentist	0.01	1.04	2.26 (0.76)
	Bayesian (Noninformative)	0.02	1.12	2.67 (1.17)
Optimal Matching	Frequentist	0.09	1.32	2.23 (0.75)
	Bayesian (Noninformative)	0.02	1.07	2.28 (0.90)

PPI. In contrast, for one data set, the frequentist propensity score approach provides only one point estimate of each balance index. To compare with the interval estimate provided by the Bayesian approach, we bootstrap 1,000 data sets from the original data to obtain a 95% bootstrap confidence interval for each balance index.

RESULTS OF THE CASE STUDY

Results of the case study are presented in Table 1 as well as in Figures 2 to 8. Table 1 shows the average absolute standardized mean/proportion difference and variance ratio across all the covariates for different propensity score methods as well as the treatment

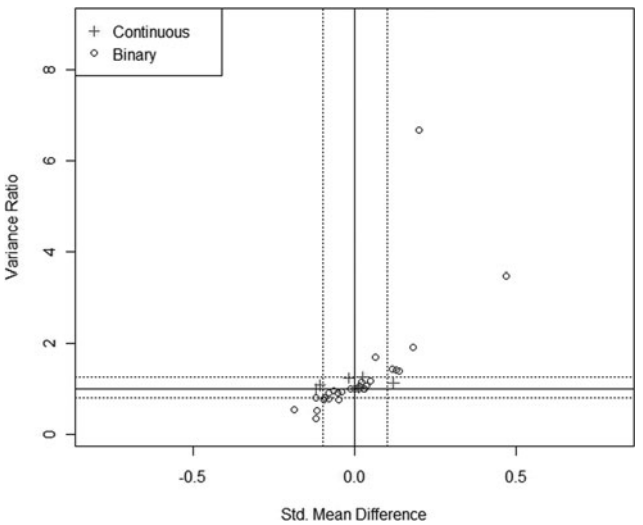

Figure 2. Initial covariate balance check in the case study.

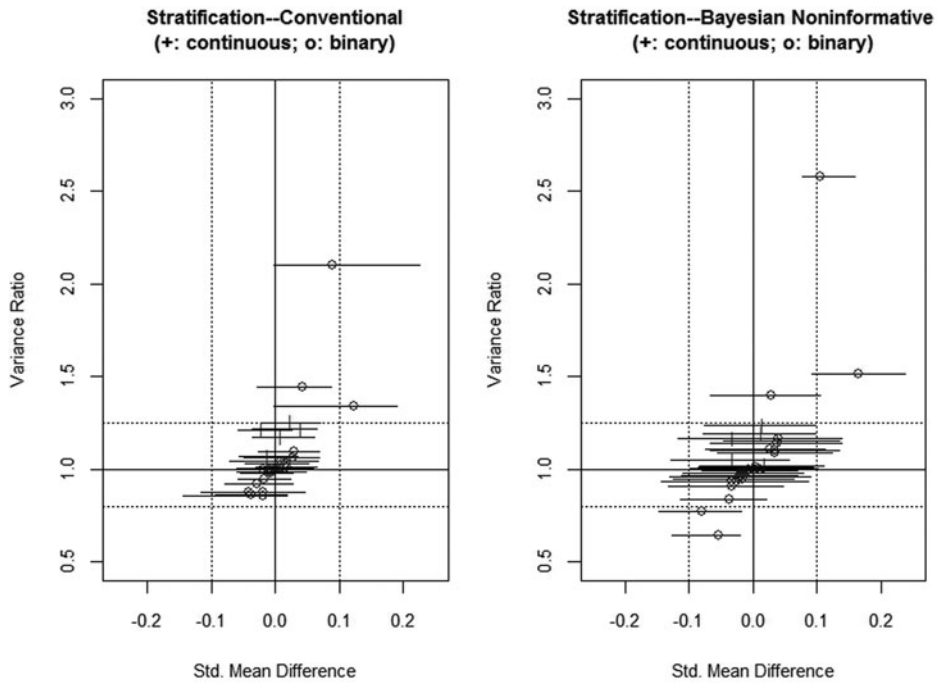

Figure 3. Standardized mean/proportion difference and corresponding 95% bootstrap/posterior probability interval for stratification in the case study.

effect estimates. In the Figures 2 to 8, the balance of each individual covariate is illustrated, where each continuous covariate is marked by a plus and each categorical level is marked by a circle. Figure 2 illustrates the initial imbalance in the sampled ECLS-K data set, and Figure 3 to Figure 8 present the interval estimates of the balance indices for propensity score stratification, weighting and optimal full matching methods.

Overall, for the point estimates of Cohen's d and variance ratio, the two-step Bayesian propensity score approach performs similarly compared to the frequentist propensity score methods across most of the continuous covariates and categorical levels. In terms of the average standardized mean/proportion difference and variance ratio, the frequentist propensity score methods provide slightly better covariate balance in stratification and weighting, whereas the two-step Bayesian method achieves better covariate balance under optimal full matching method. On average, both frequentist and Bayesian propensity score adjustment methods greatly reduce initial imbalance as shown in Table 1. After propensity score adjustments, the standardized mean/proportion differences are within the ideal range ($\pm$ 0.1 standard deviation) for most of the continuous/categorical covariates, and some covariates even achieve approximately zero standardized mean/proportion difference between the treatment group and control group as illustrated in Figure 3 to Figure 5. According to the balance criteria presented in Rubin (2001), variance ratios for most of the covariates in the treatment group and control group fall in the acceptable range (between 1/2 and 2) except for a few levels of categorical covariates. Many of the variance ratios are in the desirable range (between 4/5 and 5/4).

In addition to point estimates, Figure 3 to Figure 5 also show the 95% bootstrap confidence intervals (for conventional frequentist approach) or posterior probability intervals

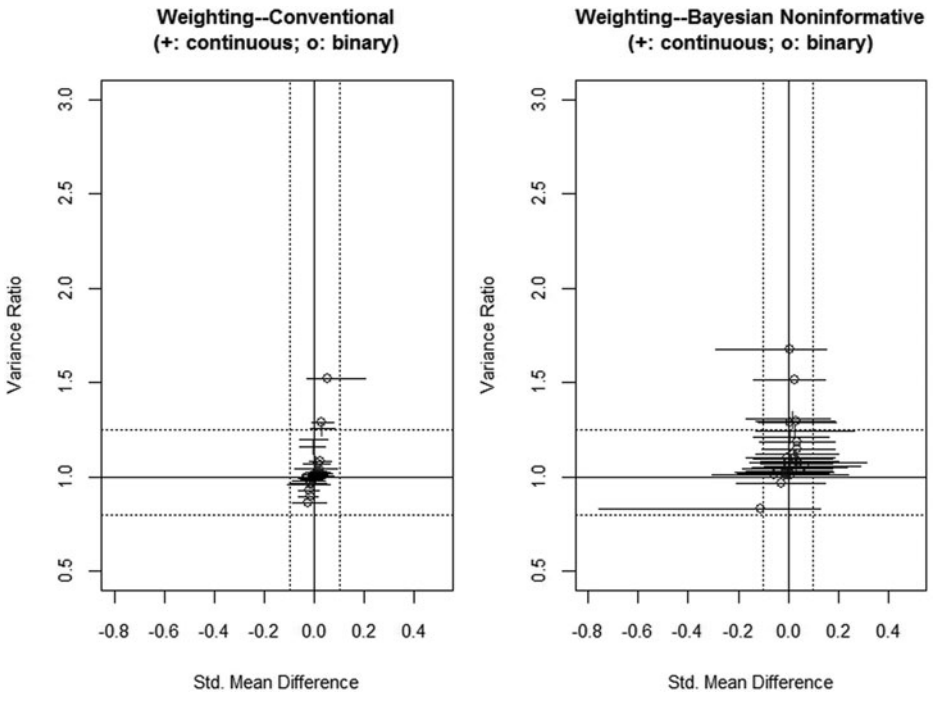

Figure 4. Standardized mean/proportion difference and corresponding 95% bootstrap/posterior probability interval for weighting in the case study.

(for Bayesian approach) of the standardized mean/proportion differences for the different propensity score methods. A 95% interval that falls in the desirable range indicates good covariate balance. For the stratification method, the *PPIs* provided by the two-step Bayesian propensity score approach are similar to the bootstrap intervals estimated by the frequentist propensity score approach. For the weighting method, a greater number of frequentist bootstrap intervals fall into the desirable range compared to the the two-step Bayesian approach, where as for the optimal full matching method, the *PPIs* produced by the two-step Bayesian approach are consistently better than the bootstrap intervals, among which a 95% bootstrap interval falls out of the desirable range completely.

Moreover, Figure 6 to Figure 8 display the 95% bootstrap confidence intervals or *PPIs* for the variance ratio of each covariate/categorical level between two treatment conditions. Similar to the point estimates, the frequentist propensity score approach provides slightly better bootstrap intervals for the stratification and weighting methods, where as the two-step Bayesian approach offers better posterior probability intervals for the variance ratio under the optimal full matching method. We note that due to some extreme bootstrap data sets, the upper bound of the variance ratio for one level of the family composition variable is approximately infinite under frequentist propensity score stratification, weighting and optimal full matching methods. Thus, no interval bar is drawn for that categorical level in Figures 6 to 8. The Bayesian propensity score approach shows unique benefits in this context because *PPI*s can be obtained simultaneously with point estimates based on the posterior distribution of the balance indices, where as bootstrap confidence intervals may sometimes be undesirable due to some extreme bootstrap samples.

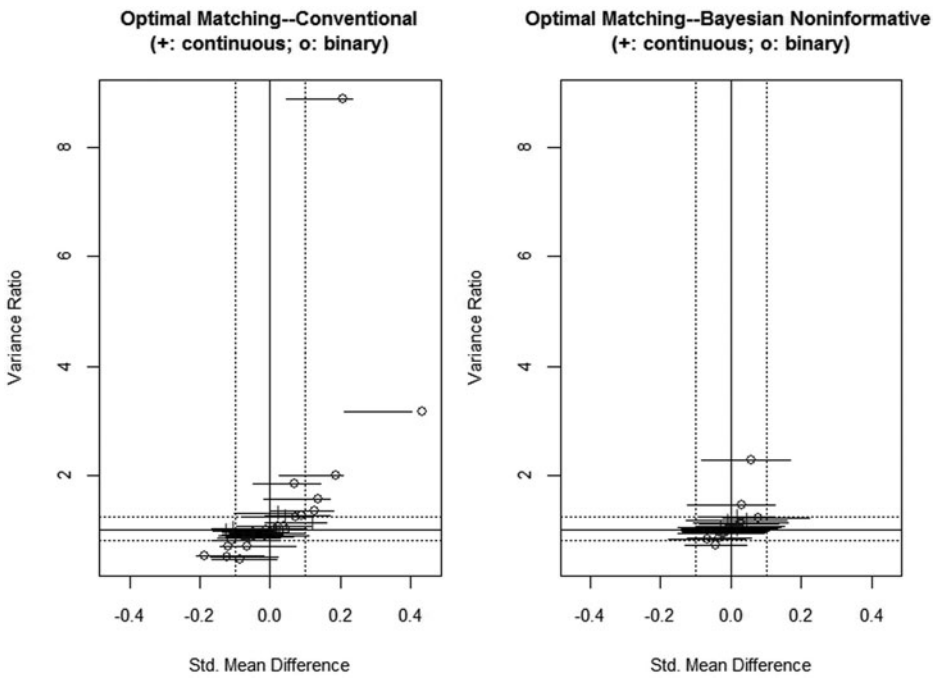

Figure 5. Standardized mean/proportion difference and corresponding 95% bootstrap/posterior probability interval for optimal full matching in the case study.

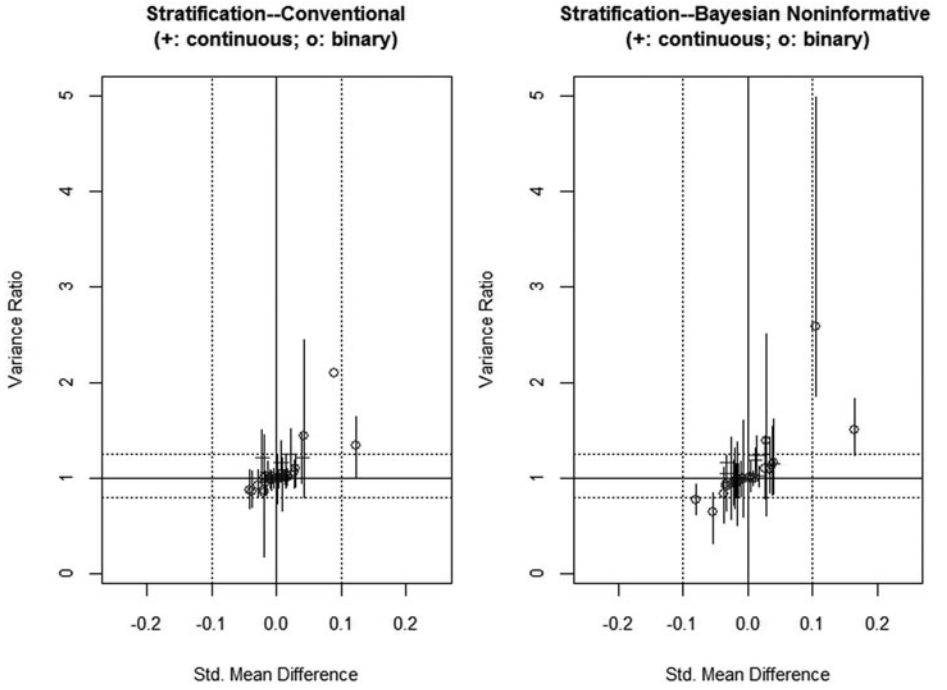

Figure 6. Variance ratio and corresponding 95% bootstrap/posterior probability interval for stratification in the case study.

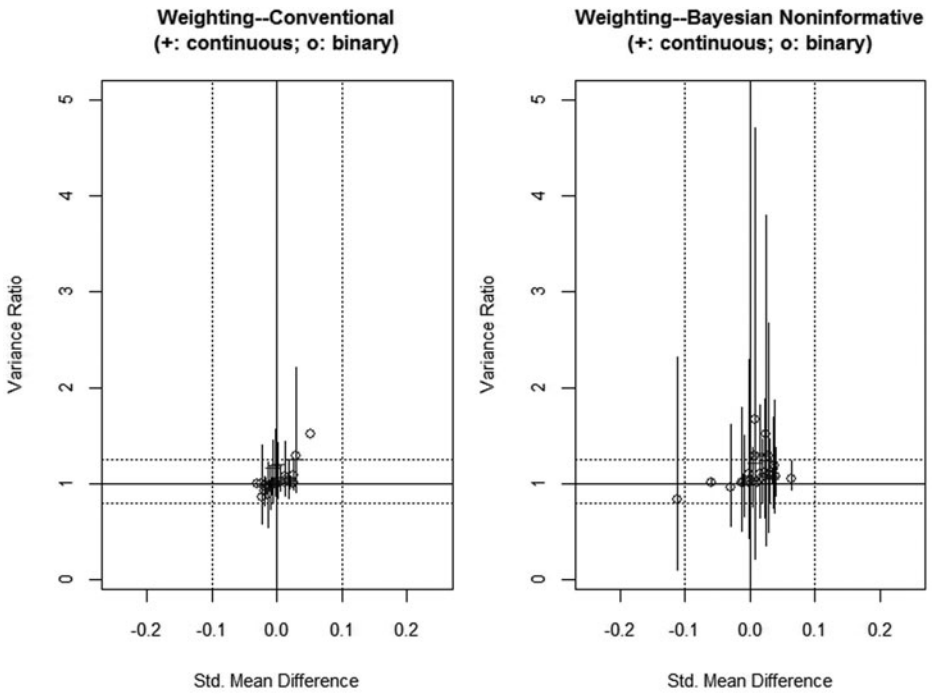

Figure 7. Variance ratio and corresponding 95% bootstrap/poterior probability interval for weighting in the case study.

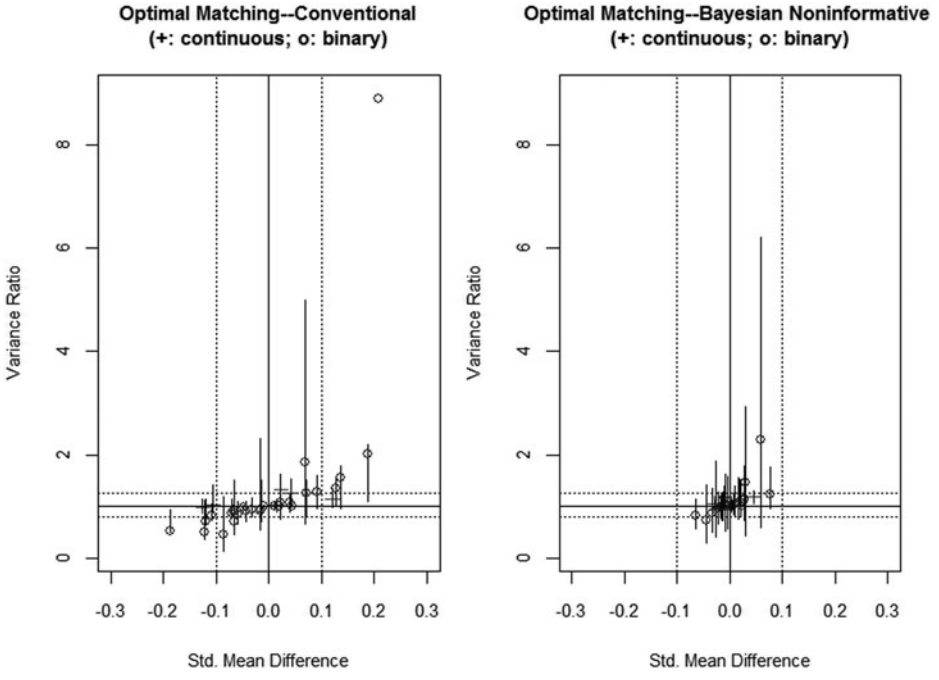

Figure 8. Variance ratio and corresponding 95% bootstrap/posterior probability interval for optimal full matching in the case study.

In terms of the consistency of performance on covariate balance and treatment effect estimates, results presented in Table 1 indicate that the two-step Bayesian propensity score approach is similar to the frequentist approach with regard to both covariate balance and treatment effect estimation for the stratification method. For the weighting method, though Bayesian and frequentist approaches provide comparable covariate balance, their treatment effect estimates differ considerably. In contrast, for the optimal full matching method, Bayesian and frequentist propensity score approaches perform similarly on the treatment effect estimation, though their covariate balance results are distinct, implying that treatment effect estimation by the optimal full matching method is relatively robust to the remaining covariate imbalance. As the true treatment effect is unknown in the case study, we further discuss the link between covariate balance and treatment effect estimates in the following simulation study.

DESIGN OF THE SIMULATION STUDY

An important feature of Bayesian techniques is the capability of incorporating prior information from expert opinion or previous research to directly address uncertainty in model parameters. Kaplan and Chen (2012) have shown that accurate prior information regarding treatment effect yields higher bias reduction and improved treatment effect estimates. Therefore, to further explore the effects of prior information on the balance properties of the two-step Bayesian propensity score approach and to further examine the link between covariate balance and treatment effect estimation, we conduct a simulation study in which 200 samples, each with size of 1,000, are bootstrapped from the observed data in the case study to mimic the real data. The same 14 covariates are adopted. Also, the balance indices and propensity score methods utilized in the simulation study are the same as those in the case study.

The parameter estimates in the propensity score model of the case study serve as true parameters to generate true propensity scores for the simulation study. Then we calculate the treatment assignment vector T by comparing the true propensity scores $e_i(x)$ to a random variable U_i generated from the $Uniform(0, 1)$ distribution, where i refers to the ith person ($i = 1, 2, \ldots, n$). We assign $T_i = 1$ if $U_i \leq e_i(x)$ and $T_i = 0$ otherwise. We then obtain the outcome variable y by a linear model with the generated treatment assignment and covariates as predictors. The true treatment effect is set as 2.29 (the estimate by covariate-adjusted regression in the case study) and the coefficients of covariates are set as 0.3 or –0.3.[2]

Three approaches were evaluated in the simulation study: (a) the frequentist propensity score approach; (b) the two-step Bayesian propensity score approach with noninformative prior, reflecting the situation of having little prior information; and (c) the Bayesian propensity score approach with informative prior, where the generating propensity score parameters are used as prior means with prior precision 100 (i.e., prior variance 0.01) to represent the situation of having strong prior information. In practice, accurate prior information can be elicited from previous research, established theory, or expert opinion. When researchers are in possession of accurate prior information, typically small prior variance (high precision) will be specified. When researchers are not certain about accuracy of the prior information, a prior with a variance (low precision) will be utilized. As with the

[2]Detailed R code is available upon request.

case study, the "bayesglm" function of the *arm* package (Gelman et al., 2011) is used for estimating parameters in the Bayesian logit model.

RESULTS OF THE SIMULATION STUDY

Before presenting the results, we note that a common nonidentifiability problem within frequentist logistic regression, pointed out by Gelman, Jauklin, Pittau, and Su (2008), occurred in the simulation study. Specifically, some covariates perfectly separated the treatment group and control group, particularly for the binary predictors. This could occur, for example, if the treatment group is composed of all male participants and the control group is composed of all female participants. A common strategy for overcoming this problem is to remove the predictors that cause the separation until the model is identified. This strategy is incorporated as the default in the "glm" function of the R "stats" package (R Development Core Team, 2011) for frequentist logistic regression. For this simulation study, Bayesian logistic regression with noninformative priors removes the same predictors that cause the nonidentifiability problem in frequentist logistic regression. In addition, for the weighting methods, we use a mild informative normal prior with prior mean 0 and prior precision 0.01 as the "noninformative" prior for the Bayesian approach in order to avoid extreme propensity score weights. However, for the two-step Bayesian propensity score approach with informative priors, nonidentifiability in the logistic regression was not a problem, even with the binary predictors that cause the separation. This indicates the important role that prior information can play in improving model identification, especially for logit models.

The average absolute standardized mean/proportion differences and average variance ratios across all the covariates as well as average treatment effect estimates across 200 bootstrapped data sets are presented in Table 2. The average balance of each individual

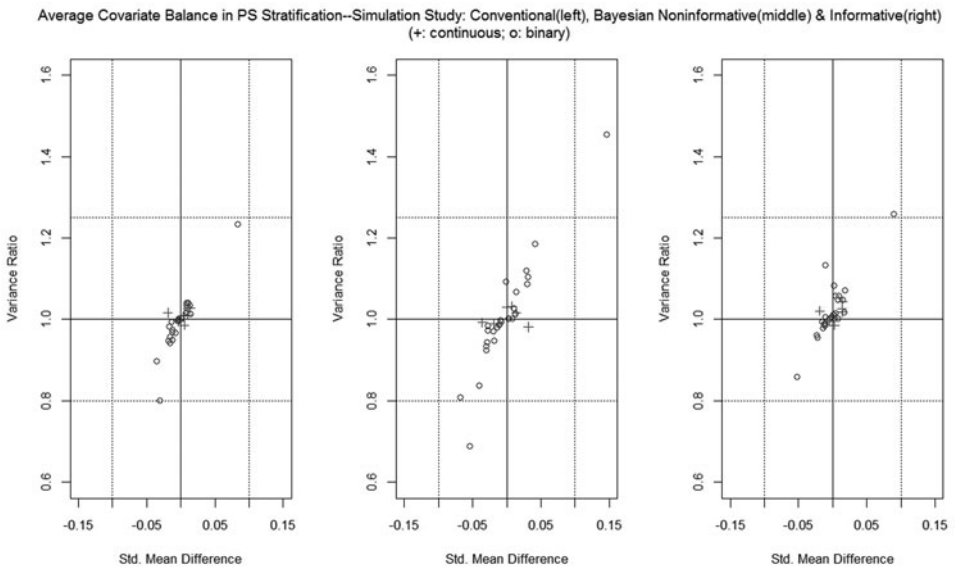

Figure 9. Average covariate balance in stratification in the simulation study.

Table 2. Average absolute standardized mean/proportion difference (Cohen's d), average variance ratio between treatment and control group across all covariates and treatment effect and standard error estimates (Trt. (SE)) in the simulation study

Method		Cohen's d	Variance Ratio	Trt. (SE)
Stratification	Frequentist	0.02	Infinite	2.39 (0.39)
	Bayesian (Noninformative)	0.03	Infinite	2.38 (0.53)
	Bayesian (Informative)	0.02	Infinite	2.92 (0.47)
Weighting	Frequentist	0.52	1.41×10^{10}	2.16 (0.40)
	Bayesian (Noninformative)	0.17	8.58×10^{3}	2.07 (0.69)
	Bayesian (Informative)	0.13	1.21	2.69 (0.49)
Optimal Matching	Frequentist	0.0033	1.03	2.24 (0.40)
	Bayesian (Noninformative)	0.0028	1.03	2.29 (0.58)
	Bayesian (Informative)	0.0028	1.03	2.77 (0.56)

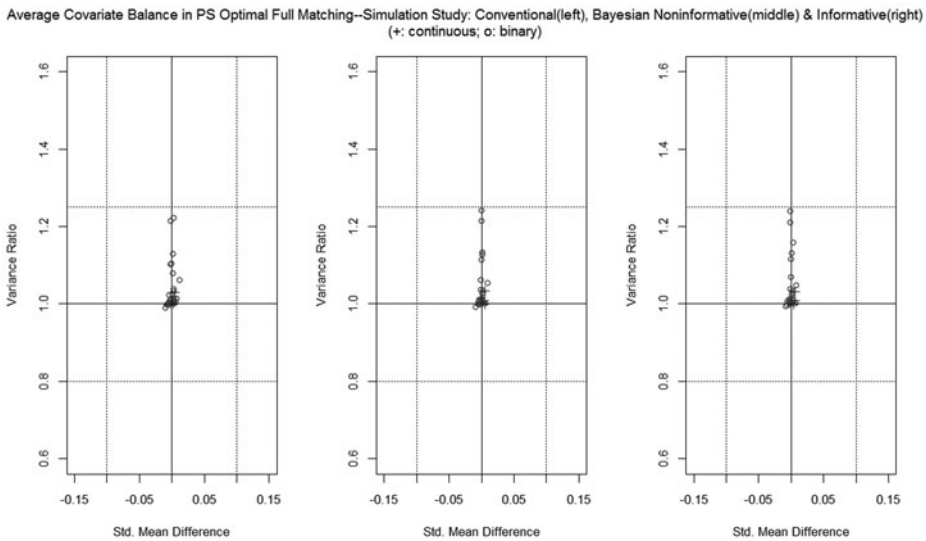

Figure 10. Average covariate balance in optimal full matching in the simulation study.

covariate across 200 bootstrapped data sets are illustrated in Figure 9 and Figure 10.[3] Overall, the frequentist propensity score approach performs similarly to the two-step

[3]The balance performance of two categorical levels are not shown in Figure 8 due to infinite variance ratios. The balance properties for the weighting method are not illustrated in the simulation study due to huge variance ratios in the frequentist propensity score approach and the two-step Bayesian approach with noninformative prior.

Bayesian propensity score approach on covariate balance for stratification and optimal full matching methods in terms of average Cohen's d as shown in Table 2. Two categorical levels out of 35 continuous covariates/categorical levels yield infinite variance ratios for the stratification method for both the frequentist and Bayesian approaches. This is mainly because the commonly used stratification method utilizes only five strata, and therefore it is hard to balance some highly imbalanced binary covariates. However, the optimal full matching method matches the control group with the treatment group much more finely and provides the best and most stable variance ratios for both the frequentist and Bayesian approaches compared to the stratification and weighting methods (the Cohen's d ranges from 0.0028 to 0.0033, and the variance ratios are all very close to 1), suggesting the overall superiority of the optimal full matching method.

For the weighting method, the two-step Bayesian propensity score approach provides better balance than the frequentist approach in terms of standardized mean/proportion differences. In regard to the variance ratio, both frequentist and two-step Bayesian propensity score approaches with noninformative prior yield very large variance ratios, which are due to some extreme propensity score estimates (close to 0 or 1) and thus extreme weights. However, by incorporating precise priors, the two-step Bayesian propensity score approach offers reasonable balance in regard to the variance ratio, indicating the need to elicit accurate prior information especially when data are extreme. Overall, the weighting method provides worse balance and treatment effect estimates, indicating that the performance of weighting can easily be affected by extreme propensity score weights.

In terms of the effects of prior information, on average, the two-step Bayesian propensity score approach with informative priors performs similarly on covariate balance compared to the same Bayesian approach but with noninformative prior. This finding indicates that although the two-step Bayesian propensity score approach with accurate prior information regarding treatment effect can provide better treatment effect estimates as shown in Kaplan and Chen (2012), the prior information with regard to propensity score model parameters have less impact on the covariate balance.

For the treatment effect estimates, frequentist approach and Bayesian approach with noninformative priors perform similarly as expected. Among all the methods, the Bayesian two-step method using optimal full matching with noninformative priors on the propensity score model offers the least biased treatment effect estimate (2.29, bias $= 0$). These results agree with the findings in Kaplan and Chen (2012) that the optimal full matching method provides more accurate treatment effect estimates for both frequentist and Bayesian propensity score approaches.

With regard to the consistency of performance on covariate balance and treatment effect estimation, Table 2 shows that there is a link between covariate balance and treatment effect estimates for the frequentist and Bayesian noninformative approaches under the optimal full matching method, where the Bayesian propensity score approach with noninformative priors offers better covariate balance and less biased treatment effect estimate compared to the frequentist approach. Also, the optimal full matching method outperforms the stratification and weighting method in terms of both covariate balance and treatment effect estimates for frequentist and Bayesian noninformative approaches. Despite achieving good covariate balance, the Bayesian propensity score approach with informative priors provides the most biased treatment effect estimates across different methods compared to the frequentist and Bayesian noninformative approaches, which may be because the same highly precise prior obtained from the case study was used for all the 200 bootstrapped data sets so that the prior information may not be reliable for some data sets. In practice, with only one data set,

precise and accurate priors for the propensity score model parameters may have a different impact on the treatment effect estimation.

CONCLUSION

The two-step Bayesian approach for propensity score analysis proposed by Kaplan and Chen (2012) maintains a fully Bayesian specification and naturally accounts for uncertainty in the propensity scores by specifying prior distributions on the propensity score model parameters while avoiding the concerns surrounding the joint modeling of Bayesian propensity score model and outcome model. This article examined the balancing properties of the two-step Bayesian propensity score approach via a case study and a real-data based simulation study for propensity score stratification, weighting and optimal full matching methods, and demonstrated a way of evaluating covariate balance in Bayesian propensity score analysis via point estimates and posterior distribution of balance indices. In addition, the link between covariate balance and treatment effect estimation was evaluated.

To summarize, results of the case study revealed that both Bayesian and frequentist propensity score approaches substantially reduced initial imbalance and their performance on covariate balance was similar in regard to the standardized mean/proportion differences and variance ratios in the treatment group and control group. Similar performance was also found with respect to the corresponding 95% bootstrap intervals and posterior probability intervals. Specifically, the frequentist propensity score approach provided slightly better covariate balance for the propensity score stratification and weighting methods, where as the two-step Bayesian approach offered slightly better covariate balance for the optimal full matching method. The link between covariate balance and treatment effect estimation varied across different propensity score methods.

Results of the simulation study showed that both Bayesian and frequentist propensity score approaches achieved good covariate balance for stratification and optimal full matching methods. The Bayesian propensity score approach with informative priors showed similar balance performance compared to the Bayesian approach with noninformative priors, indicating that the specification of the prior distribution does not greatly influence the balance properties of the two-step Bayesian approach. The optimal full matching method, on average, offered the best covariate balance and the least biased treatment effect estimates compared to stratification and weighting methods for frequentist and Bayesian noninformative propensity score approaches. Overall, the two-step Bayesian optimal full matching approach with noninformative priors performed the best in terms of both covariate balance and treatment effect estimate.

One benefit of conducting Bayesian propensity score analysis is that we can obtain a distribution of estimated propensity scores and thus a distribution of corresponding balance indices (Cohen's d and variance ratio in this article) so that the variation in balance indices can be captured in addition to the point estimates to assist in balance checking. Good balance is achieved if both the point estimates and the posterior probability intervals of the balance indices fall into the desirable range. We note that in the simulation study, some extreme bootstrap data sets led to extreme propensity score estimates and thus distorted variance ratios of covariates between treatment group and control group. However, after incorporating precise prior information, the extreme data sets were pulled toward the prior distribution and, as a result, reasonable propensity score estimates and balance indices were obtained. This is another benefit of the Bayesian propensity score approach.

One limitation we noticed in the simulation study is that the Metropolis acceptance rate in MCMC posterior sampling can be very small for the Bayesian logit model when the prior is improper and/or data are extreme. Utilizing a proper prior may improve the acceptance rate of the Metropolis sampling algorithm, which is a topic for future research. Using other computing algorithm such as the EM algorithm used in this article provides a possible solution to this problem.

In this article, we investigated the mean and variance balance for covariates in the linear response surface. When there are nonlinear relationships between covariates and outcome, covariate balance in higher order moments such as variance, squared terms and interaction terms should be given extra attention because higher order imbalance could lead to bias even if covariate mean balance has been achieved (Hill, 2008). The balance performance for the two-step Bayesian propensity score approach with nonlinear confounders is warranted for further study.

We also want to make a note here that misspecification in the propensity score model can lead to poor covariate balance and biased treatment effect estimates (Kang & Schafer, 2007), where as misspecification in the outcome model can bias the treatment effect estimates even more (Austin & Mamdani, 1993). In addition to carefully choosing covariates and specifying the functional form of the propensity score model and outcome model, Bayesian nonparametric modeling procedure, such as Bayesian Additive Regression Trees, can be an alternative, which focuses on the flexible modeling of the outcome and is shown to produce more efficient estimates in the nonlinear settings (Hill, 2011).

To conclude, this article evaluated the balancing properties of the two-step Bayesian propensity score approach and linked it with the treatment effect estimation. Consistent with the findings of Kaplan and Chen (2012), we find that the two-step Bayesian propensity approach under optimal full matching is recommended in terms of both establishing covariate balance and reducing bias in the estimated treatment effect.

ACKNOWLEDGMENTS

We acknowledge Peter Steiner for providing original R code for covariate balance within frequentist-based propensity score methods.

FUNDING

The research reported in this article was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110001 to The University of Wisconsin–Madison. The opinions expressed are those of the authors and do not necessarily represent views of the Institute or the U.S. Department of Education.

References

An, W. H. (2010). Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, *40*, 151–189.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*, 2037–2049.

Austin, P. C., Mamdani, M. M. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*, 1231–1236.

Austin, P. C., Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, *25*, 2084–2106.

Ayanian, J. Z., Landrum, M. B., Guadagnoli, E., Gaccione, P. (2002). Specialty of ambulatory care physicians and mortality among elderly patients after myocardial infarction. *New England Journal of Medicine*, *347*, 1678–1686.

Cochran, W. G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, *24*, 205–213.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, *2nd edition.* Hillsdale NJ: Erlbaum.

De Finetti, B. (1974). *Theory of Probability, Vols. 1 and 2*.New YorkNY: Wiley and Sons.

Dehejia, R. H., Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Dehejia, R. H., Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, *84*, 151–161.

Foster, E. M. (2003). Propensity Score Matching An Illustrative Analysis of Dose Response. *Medical Care*, *41*, 1183–1192.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*, 1360–1383.

Gelman, A., Su, Y. S., Yajima, M., Hill, J. L., Pittau, M. G., Kerman, J., & Zheng, T. (2011). arm: Data Analysis Using Regression and Multilevel/Hierarchical Modelsm, arm: Data analysis using regression and multilevel/hierarchical models (R package version 1.4–13) [Computer software mannual]. Availabel from http://CRAN.R-project.org/package=arm

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Ed.) (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall/CRC.

Guo, S., Barth, R. P., Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, *28*, 357–383.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, *99*, 609–618.

Hansen, B. B. (2008). The essential role of balance tests in propenisty-matched observational studies: Comments on "A Critical Appraisal of Propensity-Score Matching in the Medical Literature Between 1996 and 2003" by Peter Austin. *Statistics in Medicine*, *27*, 2050–2054.

Hansen, B. B., Klopfer, S. O. (2006). Optimal full matching and related designs via network flow. *Journal of Computational and Graphical Statistics*, *15*, 609–627.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity Score Techniques and the Assessment of Measured Covariate Balance to Test Causal Associations in Psychological Research. *Psychological Methods*, *15*, 234–249.

Hill, J. L. (2008). Discussion of research using propensity-score matching: Comments on "A Critical Appraisal of Propensity-Score Matching in the Medical Literature Between 1996 and 2003" by Peter Austin. *Statistics in Medicine*, *27*, 2055–2061.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*, 217–240.

Hirano, K., Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, *2*, 259–278.

Hirano, K., Imbens, G. W., Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1169–1189.

Horvitz, D. G., Thompson, D. J. (1952). A generalization of sampling without replacement fiom a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.

Hoshino, T. (2008). A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis*, *52*, 1413–1429.

Imai, K., King, G., Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, *171*, 481–502.

Kang, J., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 523–539.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York: The Guilford Press.

Kaplan, D., & Chen, J. (2012). A two-step Bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, *77*, 581–609.

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, M. J., Berger, K., Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, *163*, 262–270.

Leow, C., Marcus, S., Zanutto, E., Boruch, R. (2004). Effects of advanced course-taking on math and science achievement: Addressing selection bias using propensity scores. *American Journal of Evaluation*, *25*, 461–478.

Lunceford, J., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study.*Statistics in Medicine*, *23*, 2937–2960.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., Neuenschwander, B. (2009). Combining mcmc with "sequential" PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, *36*, 19–38.

Martin, A. D., Quinn, K. M., & Park, J. H. (2010). Markov chain Monte Carlo (MCMC) package. Retrived form http://mcmcpack.wustl.edu/

McCaffrey, D. F., Ridgeway, G., Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403–425.

McCandless, L. C., Douglas, I. J., Evans, S. J., & Smeeth, L. (2010). Cutting feedback in Bayesian regression adjustment for the propensity score.*The International Journal of Biostatistics*, *6*, Article 16.

McCandless, L. C., Gustafson, P., Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, *28*, 94–112.

McCandless, L. C., Gustafson, P., Austin, P. C., Levy, A. R. (2009). Covariate balance in a Bayesian propensity score analysis of beta blocker therapy in heart failure parients. *Epidemiologic Perspectives & Innovations*, *6*, 5–15.

NCES., (2001). *Early childhood longitudinal study: Kindergarten class of 1998-99: Base year public-use data files user's manual* (Tech. Rep. No. NCES 2001-029). Washington DC: U.S. Department of Education

R Development Core Team., (2011). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*, 387–394.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024–1032.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*, 318–328.

Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics*, *2*, 463–472.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*, 169–188.

Schafer, J. L., & Kang, J. (2008). Average causal effects from non-randomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*, 1334–1356.

Steiner, P. M., & Cook, D. L. (2013). Matching and propensity scores. In T. D. Little (Ed.), *The Oxford handbook of quantitative mqthods*. New York: Oxford University Press.

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of "A Critical Appraisal of Propensity-Score Matching in the Medical Literature Between 1996 and 2003." by Peter Austin.*Statistics in Medicine*, *27*, 2062–2065.

Swanson, J. M., Hinshaw, S. P., Arnold, E., Gibbons, R. D., Marcus, S., Hur, K., ... Wigal, T. (2007). Secondary evaluations of MTA 36-month outcomes: Propensity score and growth mixture model analyses. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*, 1003–1014.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67.

Wooldridge, J. M. (2002). Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, *1*, 117–139.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., & Dominici, F. (2013). Model feedback in bayesian propensity score estimation. *Biometrics*, *69*, 263–273.